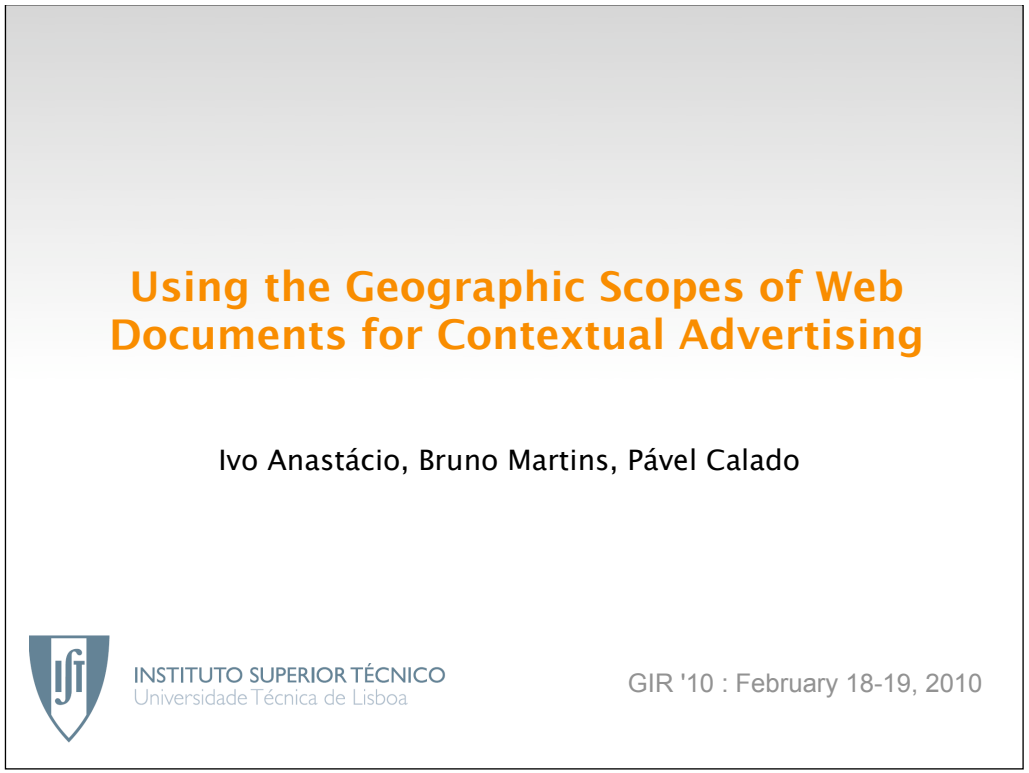


Hello my name is Ivo Anastácio and I will be presenting a work on how online advertising systems can explore the geographic scopes of web documents in order to present more meaningful advertisements to the user.

Motivation

- Geographic information is pervasive on the Web, however, little has been done on how to use such information to improve real-world search applications.
- Contextual Advertising is an example of an Information Retrieval problem that could benefit from the use of geographic knowledge.
- Study the current challenges to the development of GIR-based applications.

2

Despite the fact that current commercial search systems are currently quite good at recognizing the most relevant documents to a given theme, little has been done in order to effectively handle queries with geographic criteria.

This assumes even greater importance when one knows that geographic information is pervasive over the Internet.

Contextual advertising systems are just another type of search application, where the goal is to find the most relevant ads for a given Web page. As such, it could also benefit from better reasoning on the geographic context of the given page. Contextual advertising is the financial backbone of today's Web and an emerging research area.

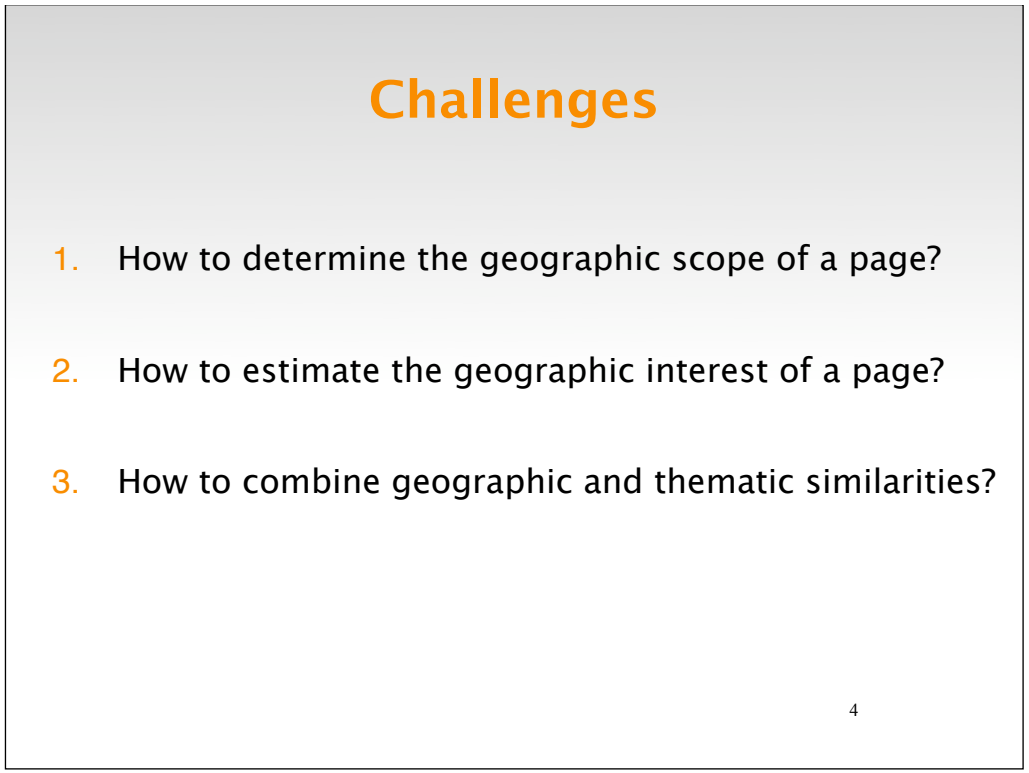
With this in mind, we wanted to develop a study that focused on the challenges of incorporating GIR in real world applications.

Motivation

- Geo-targeting based on the user-location or **geographic scope**.

Lets see a motivational example. This page describes several restaurants in the city of Florence, Italy, and on the right it has the ads Google considered interesting for the user. Based on the theme of the page and the viewer phisical location.

And the ads are, as expected, about restaurants. But in Portugal. Showing how current systems ignore the geographic scope of the documents. Clearly undermining the relevance of ads to the user, which most likely would prefer to see ads for restaurants in Florence.



Permitir que os sistemas actuais de publicidade contextualizada considerem esse âmbito geográfico não é trivial e implica principalmente responder a 3 questões.

Qual o âmbito geográfico da página em questão, uma página pode conter multiplas referencias geográficas, mas nem todas têm a mesma importância.

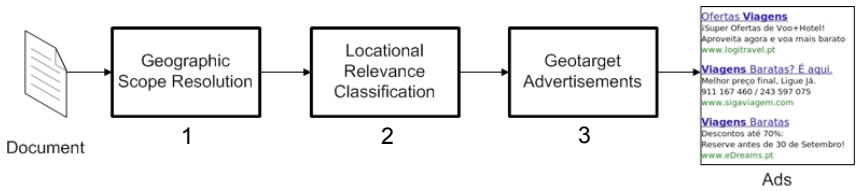
Como determinar a relevância geográfica de uma página, ou seja, mesmo que ocorram referências geográficas, estas podem não ser importantes o suficiente para influenciar os resultados.

E por fim, como combinar a similaridade temática e geográfica entre uma página e um anúncio, de modo a obter um único valor de similaridade.

Notem que estes problemas não são especificos a publicidade contextualizada, podendo influenciar outras áreas de recuperação de informação.

Objectives

- Use a combination of thematic and geographic similarities to improve the retrieval results for search applications like contextual advertising.
- Develop a prototype system capable of geo-targeting textual advertisements based on the thematic and geographic content of Web pages.

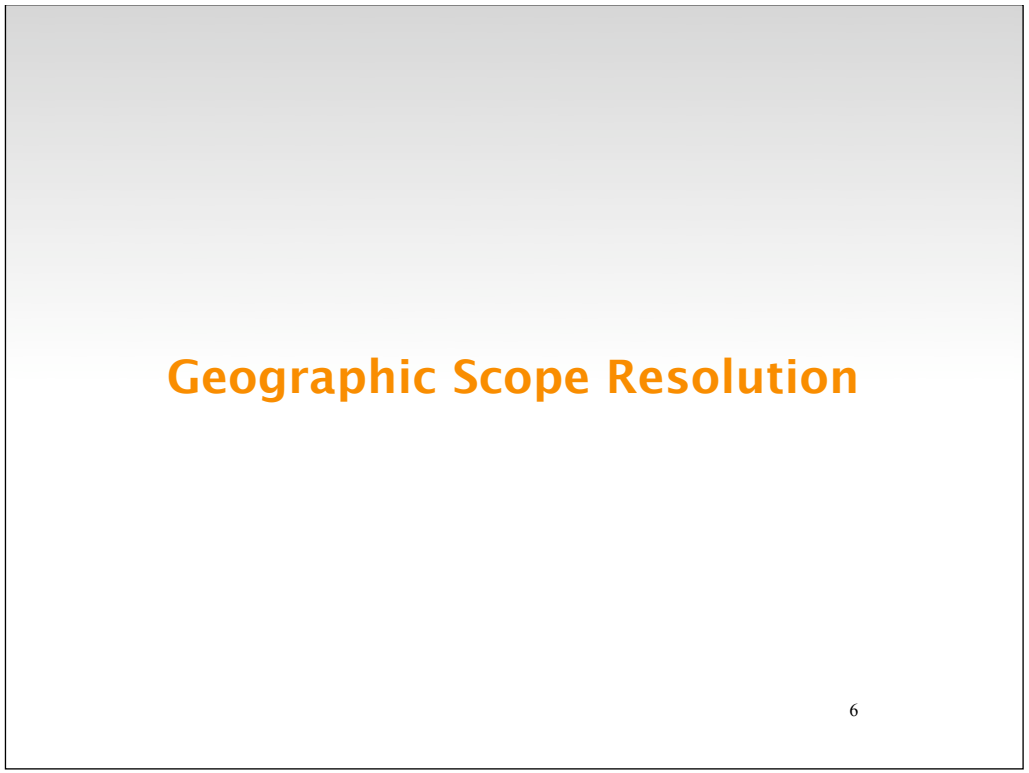


Chegamos assim à tese desta dissertação, que defende ser possivel melhorar a relevância dos anúncios retornados por sistemas de publicidade contextualizada, através da combinação da similaridade temática com a similaridade geográfica.

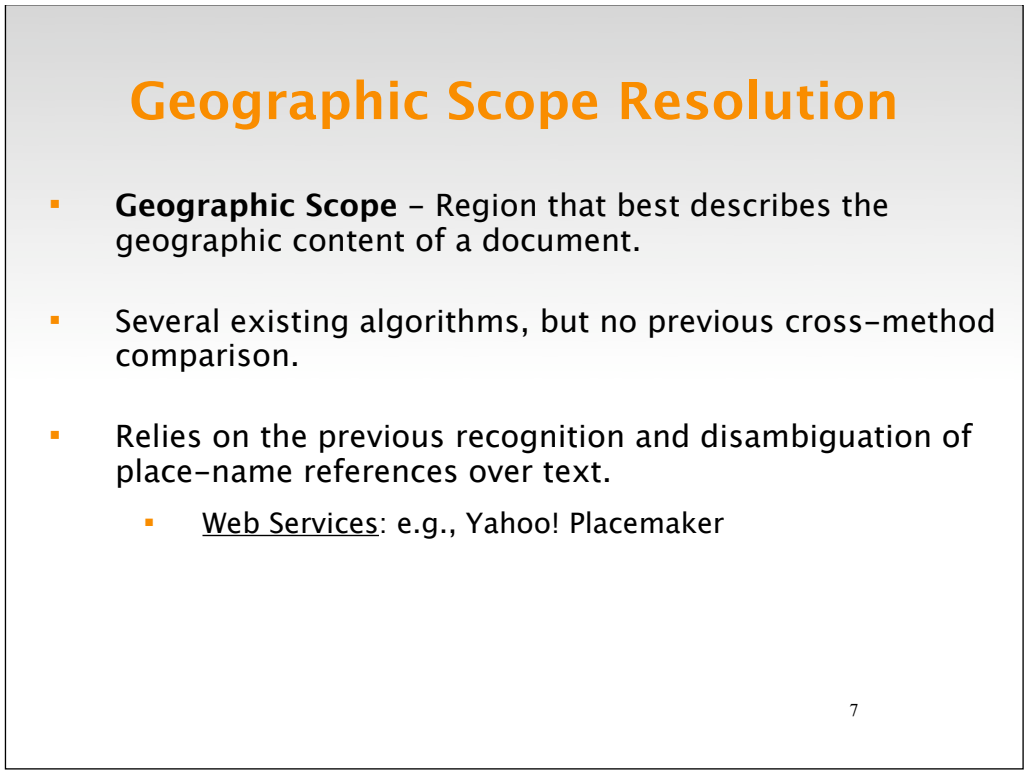
Para demonstrar experimentalmente a validade desta tese foi desenvolvido um protótipo que visa dotar um sistema de recuperação de informação tradicional com os módulos necessários para considerar a similaridade geográfica.

Nesta protótipo, existe um pipeline em que cada módulo responde a um dos desafios introduzidos no slide anterior.

Os slides seguintes descrevem cada um em maior detalhe.



Começamos pela tarefa de determinar o âmbito geográfico da página.



Um âmbito geográfico pode ser definido como a região que melhor descreve o conteúdo geográfico do documento. Isto pode implicar escolher um dos locais mencionados, ou então escolher um local que contenha alguns ou todos os locais referidos.

Para este problema já existem várias propostas, no entanto, estas não apresentam dados suficientes que permitam uma comparação directa entre elas. Havendo por isso a necessidade de se realizar um estudo comparativo de modo a escolher o método mais adequado. Foi esse o trabalho desenvolvido nesta fase.

De salientar que os algoritmos assumem que previamente os locais presentes no texto foram identificados e desambiguados. Existem já alguns Web services que oferecem essa funcionalidade, ao longo deste trabalho utilizou-se o Yahoo! Placemaker.

Geographic Scope Resolution

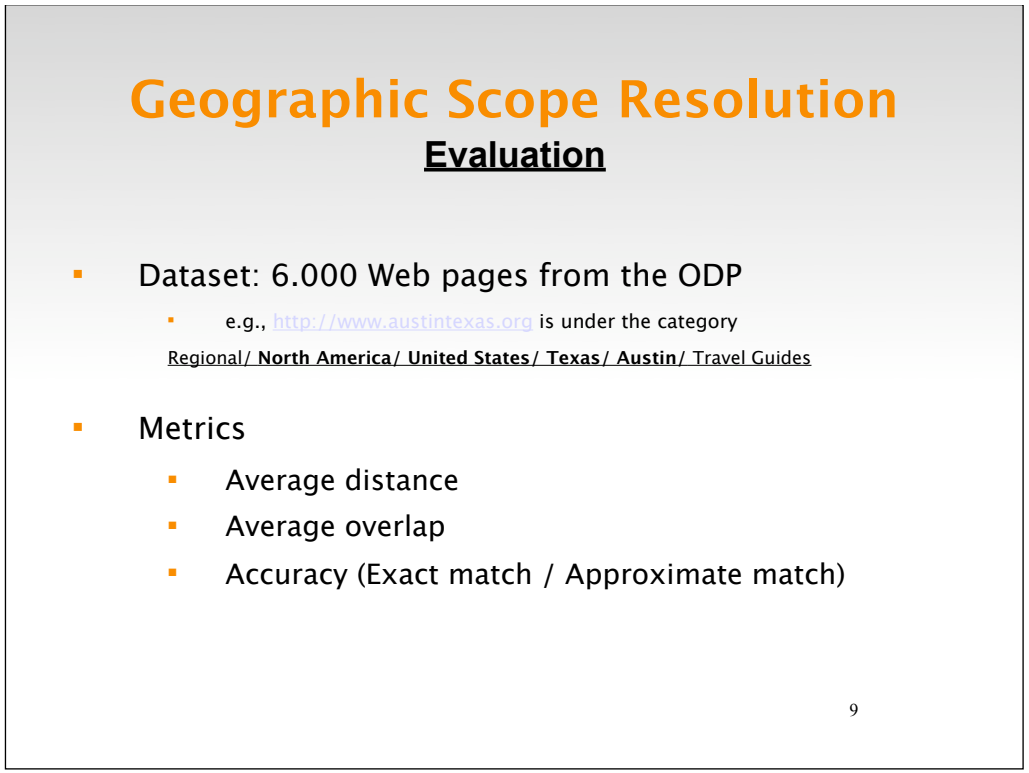
- Considered algorithms:
  - Web-a-Where
  - GraphRank
  - GIPSY
  - Yahoo! PlaceMaker
  - Baselines:
    - Most-frequent location
    - Cover Area
    - Cover area without outliers

Estes são os algoritmos considerados no estudo comparativo.

Implementei os 3 mais conhecidos.

Utilizei o Yahoo PlaceMaker, um Web Service que também oferece esta funcionalidade, mas do qual não se sabe detalhes e é testado como uma caixa negra.

E os 3 últimos são simples baselines que foram implementadas para aferir os ganhos de utilizar os métodos mais complexos.



Para avaliar os algoritmos apresentados foram utilizadas paginas do ODP, o maior directório da internet, e onde milhares de páginas se encontram manualmente categorizadas não só pela temática, mas, também em relação À geografia.

Neste exemplo vemos que esta páginas possui esta categorização que contém uma parte temática e uma parte geográfica.

A avaliação consistiu em comparar os resultados fornecidos pelos algoritmos com a categorização geográfica do ODP.

Os algoritmos foram avaliados não só em função das respostas correctas, mas também em função do erro das suas respostas. Para isso, em relação ao local correcto, mediu-se a distância média, a sobreposição média, e também as respostas correctas assumindo uma certa tolerância.

Geographic Scope Resolution

Results

Algorithm	Level	Average Distance (Km)	Average Overlap	Accuracy Distance=0	Accuracy Distance<100Km	Accuracy Overlap>0.75
GIPSY	Country	2986	0.07	0.07	0.07	0.07
	State	**442**	0.22	0.19	0.41	0.21
	City	**398**	0.37	0.16	**0.81**	0.32
	All	1275	0.22	0.14	0.43	0.2
Web-a-Where	Country	1336	0.59	0.54	0.54	0.54
	State	855	0.51	0.5	0.55	0.5
	City	704	0.42	**0.39**	0.58	0.4
	All	**959**	**0.51**	**0.48**	0.56	**0.48**
GraphRank	Country	1048	0.64	**0.61**	**0.61**	**0.61**
	State	925	0.52	0.51	0.55	0.51
	City	1281	0.34	0.33	0.47	0.34
	All	1085	0.5	**0.48**	0.54	**0.48**
Most Frequent	Country	2250	0.36	0.35	0.35	0.35
	State	501	**0.54**	**0.52**	**0.63**	**0.53**
	City	549	**0.47**	0.24	0.74	**0.45**
	All	1100	0.46	0.37	**0.57**	0.45
Covering Area	Country	2190	0.47	0	0.3	0.31
	State	3158	0.23	0	0.21	0.18
	City	2632	0.05	0	0.13	0.05
	All	2660	0.25	0	0.21	0.18
Non-outliers	Country	1523	0.57	0.45	0.5	0.55
	State	1838	0.38	0.24	0.39	0.36
	City	1872	0.12	0.02	0.28	0.1
	All	1744	0.35	0.24	0.39	0.34
Placemaker Admin.	Country	**774**	**0.71**	**0.61**	**0.61**	**0.61**
	State	1173	0.44	0.42	0.46	0.43
	City	1125	0.12	0.05	0.28	0.1
	All	1033	0.42	0.36	0.45	0.38

Estes são os resultados condensados para este estudo, na dissetação são apresentados resultados mais detalhados.

O que importa salientar aqui é que o Web-a-Where produz os melhores resultados na maioria das situações, o GraphRank também com resultados muito semelhantes ao longo de todas as métricas.

Notem no entanto que só em sensivelmente metade dos casos é que estes algoritmos acertam no âmbito.

No entanto, a baseline que escolhe o local mais frequente obteve resultados bastante competitivos, superando mesmo todos os outros para algumas métricas.

Locational Relevance Classification

Passando à 2ª tarefa, determinar se vale a pena considerar o contexto geográfico da página.

Como exemplo vejamos estas 3 páginas, contendo diferentes níveis de relevância geográfica.

A 1ª página exibe informação de cinema, sem qualquer referência geográfica no seu conteúdo. Neste caso, assume-se que não possui relevância geográfica.

A 2ª página, descreve um concerto em Lisboa, mas muitas pessoas podem visitar a página somente porque gostam da banda, independentemente de onde foi o concerto. Neste caso assume-se que existe alguma relevância geográfica.

A 3ª página é uma critica a um restaurante em Lisboa, sendo por isso maioritariamente interessante para pessoas em Lisboa, ou a planear ir a Lisboa. Neste caso assume-se que a página possui uma elevada relevância geográfica.

Locational Relevance Classification

- **Locational Relevance** – Estimate for the importance of a Web page's geographic content.
- Proposal:
  - Classify documents as **global** or **local** using a SVM
  - Use the probability estimates as the locational relevance
  - Consider two types of features:
    - Thematic (e.g., TF-IDF of all words)
    - Geographic (e.g., #locations, area of geographic scope)

13

Pode assim definir-se relevância geográfica como uma estimativa da importância do conteúdo geográfico de uma página para o utilizador.

Para determinar essa estimativa é proposto seguinte método.

Realizar uma classificação dos documentos em local ou global. Depois usar os valores de confiança com que o classificador efectuou a classificação como sendo a relevância geográfica.

Quanto maior fôr a confiança de uma classificação na classe local, maior é a relevância geográfica desse documento e vice versa.

É ainda proposto que de uma forma complementar ao texto, se utilizem features especificas ao dominio da geografia. Como o numero de locais mencionados, ou a área do âmbito geográfico.

Locational Relevance Classification

- Textual features
  - TF-IDF for all terms occurring in the document, plus all terms selected by the Yahoo! Term Extraction as the most important in the text.
- Simple Locative features
  - e.g., Total number of recognized locations.
  - e.g., Number of unique locations, grouped by city.
- High-level Locative features
  - e.g., Area for the geographic scope of the document, computed with the Web-a-Where method
  - e.g., Confidence score assigned by Web-a-Where to the scope computed for the document;

Locational Relevance Classification

Evaluation

- Dataset: 8.000 Web pages from the ODP
  - 4.000 Global
  - 4.000 Local
- 10-fold cross-validation
- Metrics:
  - Precision
  - Recall
  - F1
  - Accuracy

15

Para treinar e avaliar o classificador foram novamente utilizadas páginas do ODP. Páginas que estejam categorizadas ao nível de cidade e estado pertencem À classe local, enquanto que páginas categorizadas ao nível do país, ou sem uma categorização geográfica pertencem À classe global.

A ideia é que quanto mais especifico fôr o âmbito, maior é a relevância geográfico.

Procedeu-se a uma avaliação cruzada e avaliaram-se as medidas tradicionais, sendo que a mais importante é a accuracy, que indica de todos os documentos avaliados, que percentagem foi correctamente classificada.

Locational Relevance Classification

Results

	Recall		Precision		F-Measure			
	Local	Global	Local	Global	Local	Global	Error	Accuracy
Text	0.81	0.83	0.82	0.81	0.82	0.82	18.4	81.6
Simple Locative	0.92	0.73	0.78	0.9	0.85	0.81	17.1	82.9
High Level Locative	0.75	0.67	0.7	0.73	0.72	0.7	28.8	71.2
All Locative	0.82	0.79	0.8	0.81	0.81	0.8	19.5	80.5
Text + Best Locative	0.92	0.89	0.9	0.92	0.91	0.91	9.3	**90.7**

- Simple locative features are the most discriminative.
- Combining textual and geographic features results in improvements.

16

Mais uma vez, estes são os resultados condensados, mais combinações de features encontram-se descritas na dissertação.

O Importante a reter é que é possível distinguir entre os documentos da classe local e global com taxas de sucesso bastante boas, na casa dos 90%.

Que as features especificas à geografia do documento são as mais descriminativas.

Por fim importa salientar que realmente uma combinação de features textuais e features geográficas é a configuração que permite obter os melhores resultados.

Avançamos para a última tarefa, determinar os anúncios mais relevantes, tendo também em consideração a geografia.

Geo-target Advertisements

- **Goal**: Obtain a single similarity value for a given <page, ad> pair, which considers thematic and geographic similarities.
- Attention: geographic similarity <> geographic relevance
- Proposal:
  - Compute thematic and geographic similarities independently.
  - Linearly combine thematic and geographic similarities.
  - Use the proposed locational relevance as a dynamic weighting scheme.
    - *f(thematic_sim, geo_sim) = (1 - loc_rel) * thematic_sim + loc_rel * geo_sim*

18

O objectivo é determinar as similaridades temática e geográfica entre paginas e anúncios e combiná-las para obter uma medida única de similaridade.

Atenção não confundir relevância geográfica com similaridade geográfica, esta é por exemplo a distância entre o âmbito geográfico da página e o âmbito geográfico definido pelo anunciante para o seu anuncio.

A proposta apresentada nesta dissertação considera que as 2 similaridades são calculadas de forma independente e mais tarde combinadas de forma linear. A ideia passa por utilizar a relevância geográfica calculada na tarefa anterior como um peso dinâmico. Assim, quanto maior fôr a relevância geográfica, maior o peso dado à similaridade geográfica.

Geo-target Advertisements

- Textual Similarity
  - PostgreSQL Full-Text search similarity function
- Geographic Similarity

$$sim(S_{page}, S_{ad}) = \frac{area(S_{page} \cap S_{ad})}{area(S_{page} \cup S_{ad})}$$

$$sim(S_{page}, S_{ad}) = \begin{cases} 1 \text{ if } S_{page} \text{ is contained in } S_{ad} \\ 1 - \frac{1+sign(D)\times(1-e^{-(\frac{D}{d\times 0.5})^2})}{2} \text{ otherwise} \end{cases}$$

Geo-target Advertisements

Evaluation

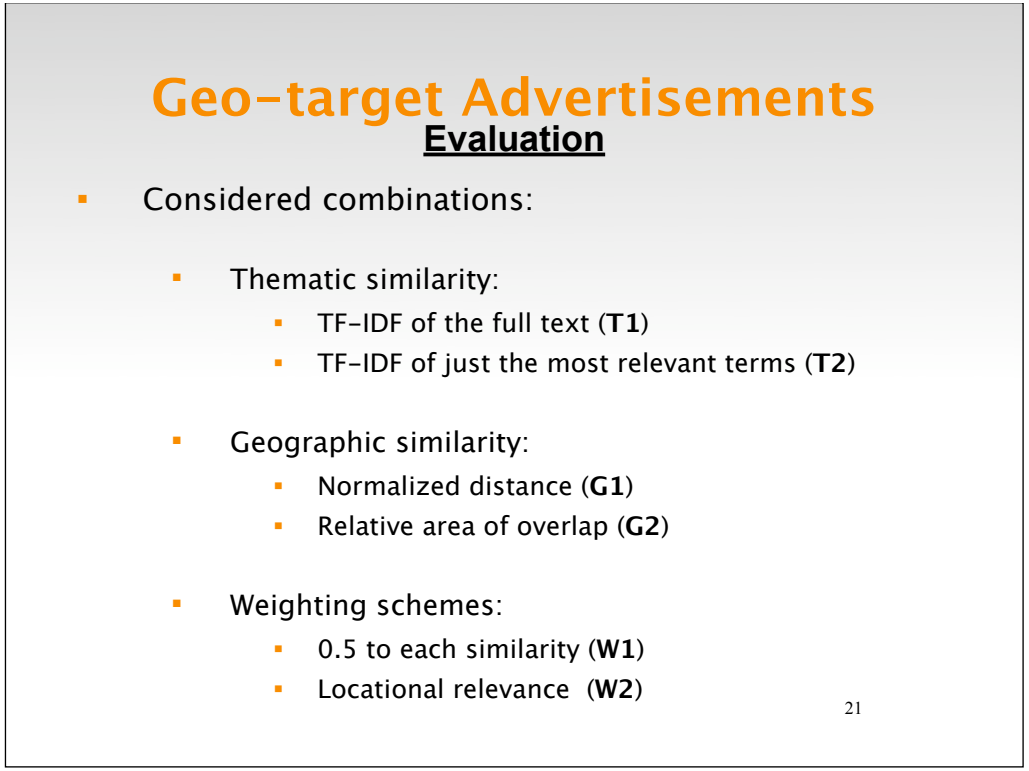
- **Local dataset**: 20 Web pages from the regional sub-sections of the ODP
  - e.g., Real Estate Guides, Travel Guides
- **Global dataset**: 20 Web pages from outside the regional section of the ODP
  - e.g., Computer Guides, Investing Guides
- **Ads dataset**: ~100.000 ads obtained from the descriptions of Business pages from the ODP
- Metrics:
  - Precision@k, k in {1, 3, 5}
  - Mean average precision (MAP)

20

Face à inexistência de colecções que pudessem ser reutilzadas para avaliar esta tarefa foram desenvolvidas colecções e julgamentos de relevância artificiais, em que se escolheu manualmente temáticas de interesse local, como guias imobiliarios. E temáticas de interesse global, como guias de computadores.

A colecção de anúncios foi gerada com base nas descrições de páginas de negócios do ODP.

Depois, manualmente foi feito um mapeamento entre as temáticas das páginas e as temáticas dos anúncios que que lhes fossem relevantes.



Testaram-se várias combinações de caracteristicas, nomeadamente:

as features usadas para a similaridade temática, usar o texto todo das páginas ou só as palavras relevantes.

Os métodos para calcular a similaridade geográfica, que são as funções actualmente consideradas como o state-of-the-art.

e o esquema de pesos com base na relevância geográfica contra um baseline que é a simples média aritmética das 2 similaridades.

Geo-target Advertisements

Results

		P@1	P@3	P@5	MAP
Local Dataset	T1	0.2	0.15	0.15	0.21
	T2	0.2	0.22	0.21	0.29
	G1	0.1	0.07	0.06	0.1
	G2	0.05	0.07	0.06	0.1
	G1T1W1	0.45	0.35	0.35	0.42
	G1T1W2	0.45	0.35	0.35	0.42
	G2T1W1	0.4	0.48	0.43	0.52
	G2T1W2	**0.45**	**0.48**	**0.45**	**0.54**
	G1T2W1	0.35	0.33	0.32	0.36
	G1T2W2	0.35	0.33	0.32	0.36
	G2T2W1	0.4	0.43	0.41	0.48
	G2T2W2	0.4	0.43	0.41	0.48
Global Dataset	T1	0.25	0.37	0.28	0.36
	T2	**0.6**	**0.43**	**0.4**	**0.58**
	G1	0	0	0	0
	G2	0	0	0	0
	G1T1W1	0.15	0.08	0.07	0.16
	G1T1W2	0.2	0.15	0.13	0.2
	G2T1W1	0.15	0.16	0.14	0.2
	G2T1W2	0.25	0.2	0.2	0.24
	G1T2W1	0.3	0.23	0.19	0.35
	G1T2W2	0.3	0.22	0.2	0.31
	G2T2W1	0.35	0.23	0.24	0.33
	G2T2W2	0.35	0.28	0.28	0.37

- The locational relevance scheme, obtained the best results for the local dataset.
- Over the global dataset, the locational relevance scheme obtained better results than the baseline.

Estes foram os resultados.

Aqui há a salientar que na colecção local, utilizar a combinação de similaridade tematica e geográfica produz os melhores resultados, com uma precisão media de 54%

Na colecção global, é importante ver que a relevância geográfica permite obter melhores resultados que o baseline.

há assim fortes indicações de que a relevância geográfica é adequada para combinar estas 2 similaridades.

Conclusions

- Contributions
  - First cross-method comparison between scope resolvers
  - First proposal for measuring the locational relevance
  - Evaluation of a new similarity combination scheme for Geographical Information Retrieval
- Results
  - Web-a-Where assigned the correct scope in 47% of the times
  - Locational relevance classification achieved an accuracy of 90%
  - A combination of geographic and thematic similarities improved the MAP in 25%

23

Concluindo, como contribuições deste trabalho, foi efectuado o 1º estudo comparativo entre métodos para atribuir âmbitos,

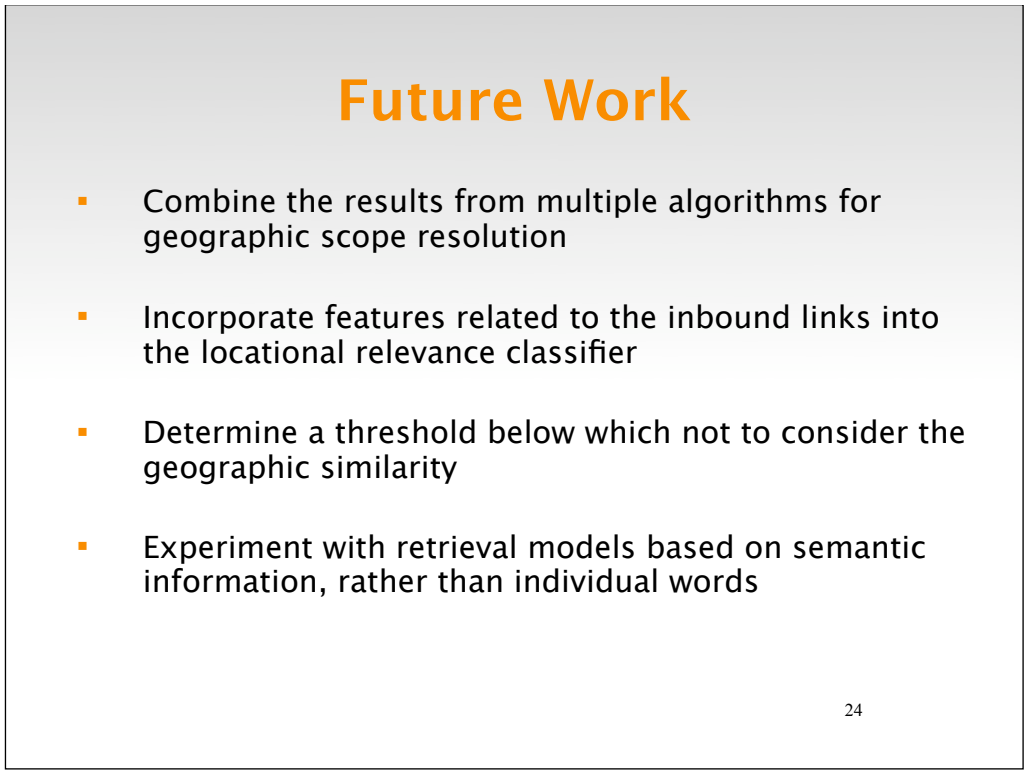
foi proposto um método para estimar a relevância geográfica de um documento

e foi testado um novo esquema para combinar similaridade temática e geográfica entre uma página e um anúncio.

Em termos de resultados, o melhor algoritmo para determinar âmbitos geográficos acerta sensivelmente em 47% dos casos.

O melhor classificador para determinar a relevância geográfica acertou na categoria correcta em 90% das páginas.

E uma combinação de similaridade temática e geográfica melhorou a precisão média em 25% face a uma abordagem puramente temática para os documentos locais.



Como trabalho futuro, existem várias oportunidades.

A comparação entre métodos para atribuir âmbitos revelou que ha bastante espaço para melhorias, pode experimentar-se a combinação de vários algoritmos, de modo a combinar os pontos fortes de cada 1.

No classificador para obter a relevância geográfica seria interessante para uma dada página incluir features contendo informação sobre os contextos geográficos das páginas que para ela apontam.

Na combinação de similaridades, seria importante determinar um limite abaixo do qual não se considerasse a similaridade geográfica.

Outro trabalho futuro seria experimentar com modelos de recuperação de informação que considerassem informação semântica, por oposição aos modelos tradicionais que consideram as palavras individualmente.

The End

Questions?